




Copyright 2012 California Institute of Technology. Government sponsorship acknowledged.

## Image Processing on the Cloud

Emily Law

Cloud Computing Workshop  
ESIP 2012 Summer Meeting  
July 14<sup>th</sup>, 2012



## Outline

CALIFORNIA INSTITUTE OF TECHNOLOGY

- Cloud computing @ JPL SDS
- Lunar images
- Challenge
- Image tiling process
- Implementations
- Analysis
- Summary

Copyright 2012 California Institute of Technology. Government sponsorship acknowledged.

2



## Science Data Systems

CALIFORNIA INSTITUTE OF TECHNOLOGY

- Cover a wide variety of domain disciplines
  - Solar system exploration, Astrophysics, Earth science, Biomedicine, etc,...
- Each has its own communities, standards and systems
- But, there is a set of common components & constraints
- Some can greatly benefit from proven cloud computing technology



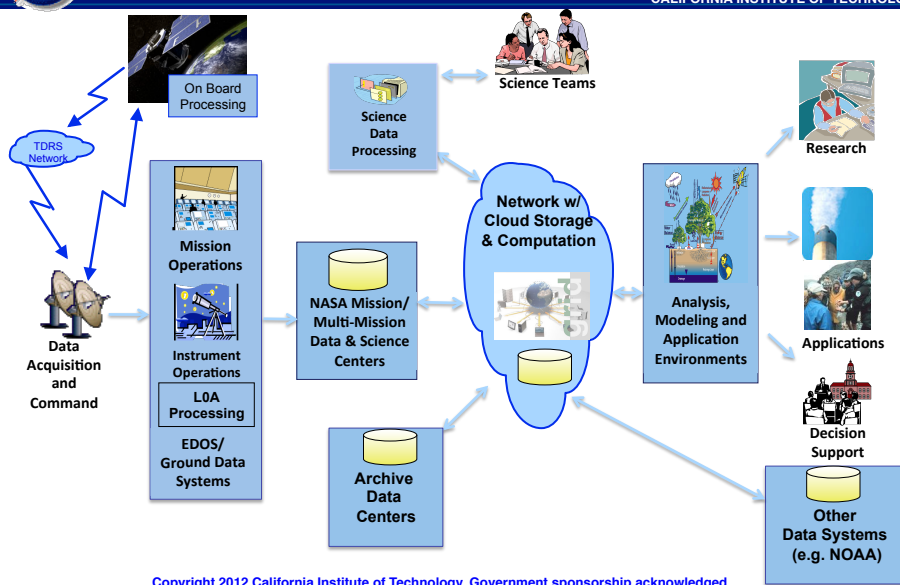
Copyright 2012 California Institute of Technology. Government sponsorship acknowledged.

3




## Earth Science Data Systems

CALIFORNIA INSTITUTE OF TECHNOLOGY



Copyright 2012 California Institute of Technology. Government sponsorship acknowledged.

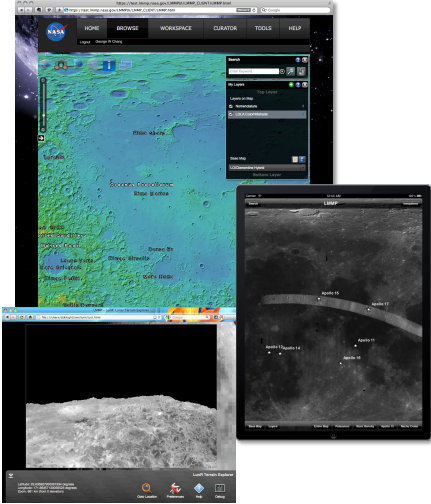
4



## Lunar Modeling and Mapping Project (LMMP)


CALIFORNIA INSTITUTE OF TECHNOLOGY

- Provides science and exploration community a suite of lunar mapping and modeling tools and products that support the lunar exploration activities
- The tools and products are made available through a common, intuitive NASA portal
- Utilizes open standards and facilitates platform and application independent access



Copyright 2012 California Institute of Technology. Government sponsorship acknowledged.


5



## Challenge

CALIFORNIA INSTITUTE OF TECHNOLOGY

- How to make these large images usable by desktop computers, mobile devices and other memory constrained products?



Copyright 2012 California Institute of Technology. Government sponsorship acknowledged.

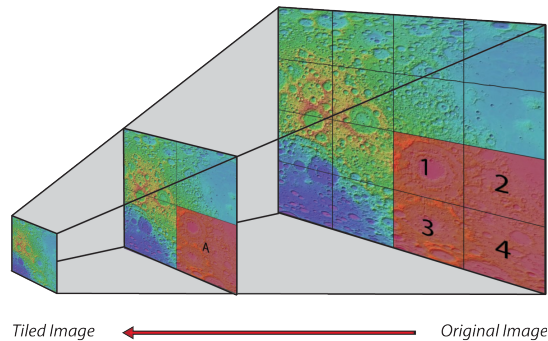
6



## Tiling Process

CALIFORNIA INSTITUTE OF TECHNOLOGY

- Divides images into small tiles
- Combines and shrinks for the next zoom level
- Iterates till the zoom level has only 1 tile



Copyright 2012 California Institute of Technology. Government sponsorship acknowledged.

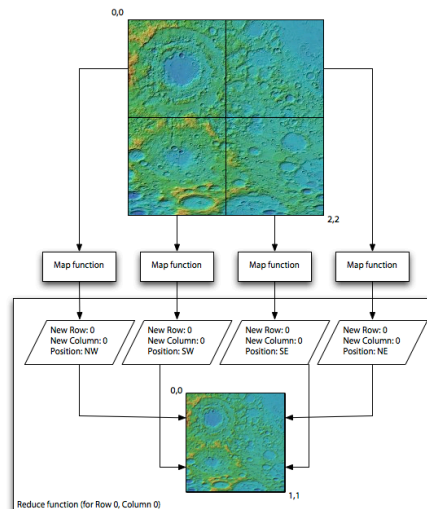
7



## Using Hadoop

CALIFORNIA INSTITUTE OF TECHNOLOGY

- Hadoop is an implementation of Google's Map-Reduce algorithm
- *Map Function* – Takes a subset of the data, performs a computation, and returns an output.
- *Reduce Function* – Consolidates outputs from the *map* function to generate another output



Copyright 2012 California Institute of Technology. Government sponsorship acknowledged.

8



## In-House Implementation

CALIFORNIA INSTITUTE OF TECHNOLOGY

- Test image, 2.77 gigabytes LRO LOLA (Lunar Orbiter Laser Altimeter) colorized digital elevation map which produced 9.1 gigabytes set of tiles
- Ran Hadoop on local machines in the lab
- 2 Sun Fire x4170 machines running dual Xeon X5570 processors with 72 GBs of RAM with a heterogeneous mix of Solaris 10 and Linux
- Performance was excellent
- Machines are costly to maintain, especially since these tasks are “bursty”

Copyright 2012 California Institute of Technology. Government sponsorship acknowledged.

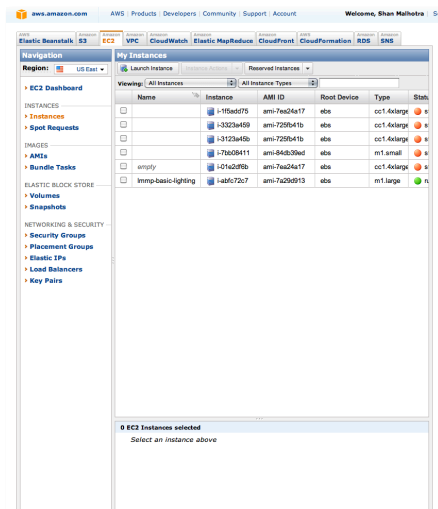
9



## Cloud Implementation Using Amazon EC2

CALIFORNIA INSTITUTE OF TECHNOLOGY

- Amazon EC2 is a cloud computing infrastructure allowing users to “rent” virtual machines
- Installed Hadoop Elastic MapReduce framework on a number of EC2 instances
- Output image files stored on Amazon S3, a cloud storage system



Copyright 2012 California Institute of Technology. Government sponsorship acknowledged.

10



## Configurations

CALIFORNIA INSTITUTE OF TECHNOLOGY

- Configuration 1 - In-House  
2x Sun Fire 4170  
72 GB RAM, 64 GB SSD Storage  
\$10K each, plus administration and infrastructure costs
- Configuration 2 - 20 EC2 “Large”  
20 EC2 Large Instances (4 Compute Units ~ 4x1GHz Xeon)  
7.5 GB RAM, 850 GB Storage  
\$0.34/instance/hour plus bandwidth
- Configuration 3 - 4 EC2 “CC”  
4 EC2 Cluster Compute Instances (33.5 Compute Units)  
23 GB RAM, 1.69 TB Storage  
\$1.60/instance/hour plus bandwidth

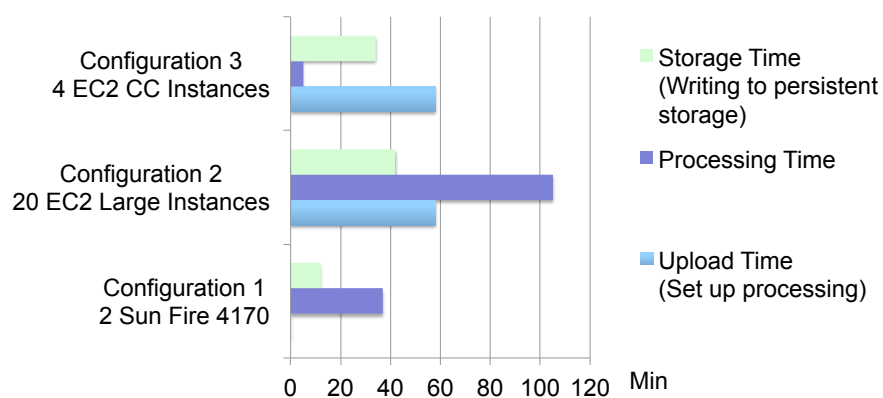
Copyright 2012 California Institute of Technology. Government sponsorship acknowledged.

11




## Performance

CALIFORNIA INSTITUTE OF TECHNOLOGY



Copyright 2012 California Institute of Technology. Government sponsorship acknowledged.

12




## Cost

CALIFORNIA INSTITUTE OF TECHNOLOGY

- In-House Implementation
  - **Total Cost: \$20K + SA + infrastructure**
- 20 EC2 Large
  - Processing:  $2h \times 20 \times \$0.34 = \$13.60$
  - Bandwidth:  $3GB \times \$0.10 = \$0.30$
  - Storage:  $10GB \times \$0.14 = \$1.40/\text{month}$
  - Total Cost: \$15.30**
- 4 EC2 CC
  - Processing:  $1h \times 4 \times \$1.60 = \$6.40$
  - Bandwidth:  $3GB \times \$0.10 = \$0.30$
  - Storage:  $10GB \times \$0.14 = \$1.40/\text{month}$
  - Total Cost: \$8.10**

Copyright 2012 California Institute of Technology. Government sponsorship acknowledged.

13



## Performance Analysis

CALIFORNIA INSTITUTE OF TECHNOLOGY

<b>In-House Implementation</b>	<b>Cloud Implementation</b>
<ul style="list-style-type: none"> <li>• Fastest overall</li> <li>• Did not need to export data to remote systems</li> <li>• Most expensive from a cost-benefit perspective</li> </ul>	<ul style="list-style-type: none"> <li>• Upload and storage time a consideration</li> <li>• Network speed between Hadoop nodes a significant consideration</li> <li>• Most cost-effective for occasional, computationally intensive jobs</li> </ul>

Copyright 2012 California Institute of Technology. Government sponsorship acknowledged.

14



## Conclusion

CALIFORNIA INSTITUTE OF TECHNOLOGY

- Hadoop framework provides a simple programmatic interface for developing distributed computing applications for problems that are parallelizable
  - Problems that required large amounts of data will depend on the interconnect speeds between nodes
- Cloud computing gives a cost-effective infrastructure to use compute capacity as needed
- In designing applications for cloud, must consider the performance of locally run machines vs. the price of cloud instances
- Security should also be considered in using public infrastructure
  - We are using a hybrid system where private data is hosted locally while public data is on the cloud

Copyright 2012 California Institute of Technology. Government sponsorship acknowledged.

15